Evaluating research portfolios, a method and a case

van den Besselaar, P.A.A.; Khalili, A.; Sandstrom, Ulf

2017

document version
Peer reviewed version

SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

Evaluating research portfolios, a method and a case[1]

Peter van den Besselaar[*], Ali Khalili[**] and Ulf Sandström[***]

[*]p.a.a.vanden.besselaar@vu.nl
Vrije Universiteit Amsterdam, Network Institute & Institute for Societal Resilience, De Boelelaan 1081, 1081 HV, Amsterdam (Netherlands)

[**]a.khalili@vu.nl
Vrije Universiteit Amsterdam, Network Institute & department of computer Science, De Boelelaan 1081, 1081 HV, Amsterdam (Netherlands)

[**]ulf.sandstrom@indek.kth.se
KTH Royal Institute of Technology, Indek ITM, SE-10044, Stockholm (Sweden)

ABSTRACT
Evaluating whether a portfolio of funded research projects (of a research council), or a portfolio of research papers (the output of a university) is relevant for science and for society required two-dimensional mapping of the project portfolio: (i) projecting the portfolio on a science map showing how the portfolio fits into and possibly shapes the research fronts, and (ii) projecting the portfolio on a map of societal challenges, showing where the portfolio links to societal problem solving or innovation. This requires evaluating in two different 'languages': a technical language relating projects to the research front, and a societal language relating the projects to societal challenges. In this paper, we demonstrate a method for doing so, using the SMS-platform. The advantage is that the method is much less dependent on subjective classifications by single experts or a small group of experts, and that it is rather user-friendly.

INTRODUCTION
Evaluating funding programs, or research output has at least two dimensions: is the portfolio adequate in (i) scientific and (ii) societal terms. A way to do this could be through a double annotation process, where project descriptions or academic papers are annotated using a knowledge base with a taxonomy for the science fields involved, and a knowledge base with a taxonomy for one or more societal challenges addressed by the portfolio. Using those knowledge bases – which are generally not an individual but a collective product – overcomes the problem that individual experts that would annotate the projects or papers are always biased and may select a biased set from the list of terms extracted from the material. This hold for technical keywords related to research fields as for technical terms relating to the societal challenges. Furthermore, annotating by experts is a time-consuming task, and therefore an automatic procedure would be helpful. The approach is based on the SMS[2] platform, the

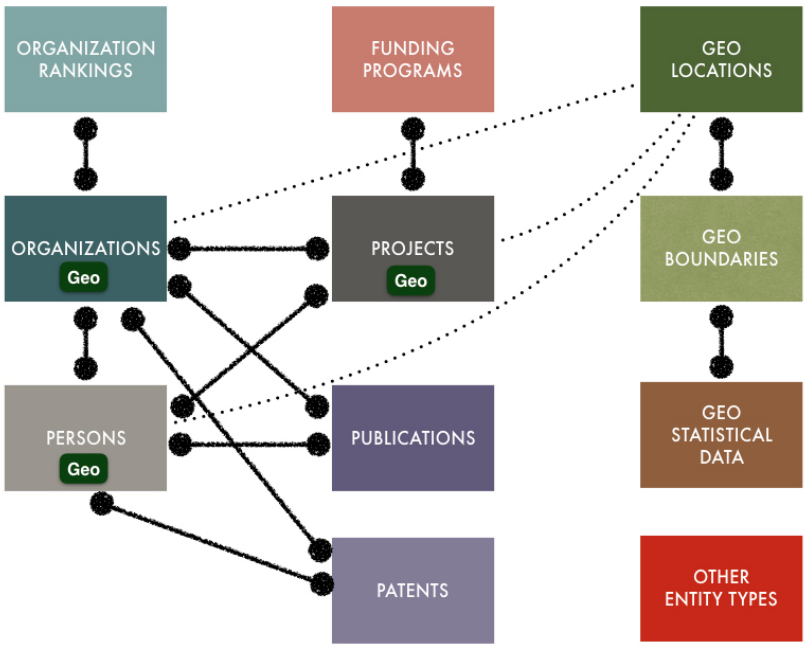
[1] This work was supported by the EC, Grant Agreement n°313082, the RISIS project.
[2] www.sms.risis.eu

technical core of the RISIS infrastructure[3], and it makes use of the increasingly rich sphere of Linked (Open) Data. We use the projects funded in H2020 as example, but a similar approach could be used for e.g. the paper production of a university – using full text or abstracts of the papers.

THE TOOLS

SMS is a 'big data' platform where 'big' is mainly used in the meaning of being 'heterogeneous'. The platform integrates a variety of datasets of different types and formats. The platform is currently focusing on data relevant for *science, technology and innovation studies*, including science policy studies and research evaluation studies. Although the platform is primarily developed to *support research* in this area, it is also *useful for evaluation exercises*, as we will illustrate. In terms of technology, the SMS platform is based on principles of semantic web[4], and linked open data[5]. The platform consists of three layers: (i) the data layer, in which data are converted into a standard format, linked with each other and included in a data store (figure 1); (ii) the service layer for enriching and harmonizing the data, and (iii) the application layer which provides a set of user interfaces on top of one or more services. In the data store, some dozens of datasets about researchers, R&D organisations, funding schemes and granted R&D projects, and R&D output (publications, patents) are linked. Additionally, the data store contains (links to) geographical and statistical data which are used for enriching the STI related data. Several of the data sets included are open data, others are private or confidential.

Figure 1. Main interlinked entity types for data integration in SMS platform



Faceted browser

A core application within SMS is the *faceted* browser, which enables the user to browse the linked data. In the browser we can reduce the data by selecting properties. E.g., one may

[3] www.risis.eu
[4] https://www.w3.org/standards/semanticweb/
[5] http://lod-cloud.net/
http://linkeddata.org/

select within the larger set of all R&D intensive organizations (about 75000) only the higher education institutions, or the hospitals, and continue with that subset only. Most figures in this paper are screenshots of the browser. The browser help to get acquainted with the data, and gives a first qualitative idea about the project portfolio, the research topics and the societal issues addressed. By selecting the relevant projects (using the relevant annotated terms – see next section), a *SPARQL query* (fig 12) is produced for retrieving the relevant data from the data store for further analysis and visualization.

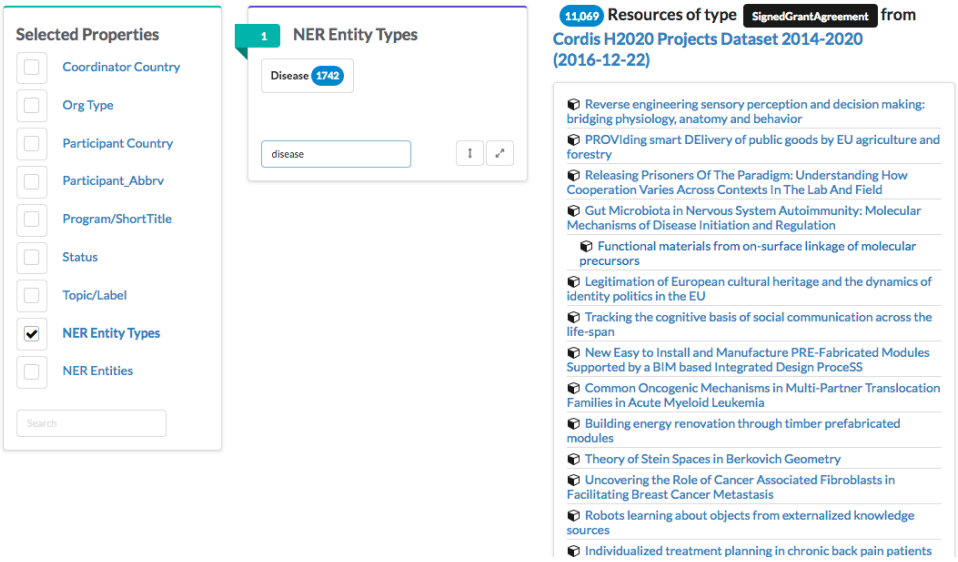
The annotation tool

Another application within SMS is the *annotation tool*, that can be used to annotate text fields using (existing) knowledge bases. Currently we deploy the *DBpedia Spotlight*[6] tool which contains a few knowledge bases, such as DBpedia, Yago and Schema.org, but we are planning to have more knowledge bases integrated, with rich concept taxonomies for different knowledge domains. The more specific the taxonomies, the better one describe texts field through these annotations. Obvious candidates for annotation are summaries/abstracts of projects and papers. The better the taxonomies, the more precisely the content of a paper portfolio or of a project portfolio can be described.

DATA AND METHODS

For evaluation both tools are useful. To illustrate this, we use the *Cordis* open dataset with H2020 projects (version December 2016). The data were downloaded from the EC website, and converted into RDF format – the standard for linked data. This enables us to inspect (in the faceted browser) and analyse the data. The browser shows the relevant characteristics of the projects, such as organizations involved, the organization type, and the program the project belongs to (figure 2). The CORDIS dataset contains among others a text summarizing the content of the projects. This is a relatively short text, but it would not be difficult at all to couple full project descriptions (e.g., all full text granted applications) to the SMS platform. It would be useful to experiment with this, and try to find out what textual information leads to the most accurate representation of the projects.[7]

Figure 2: Finding projects on 'diseases'



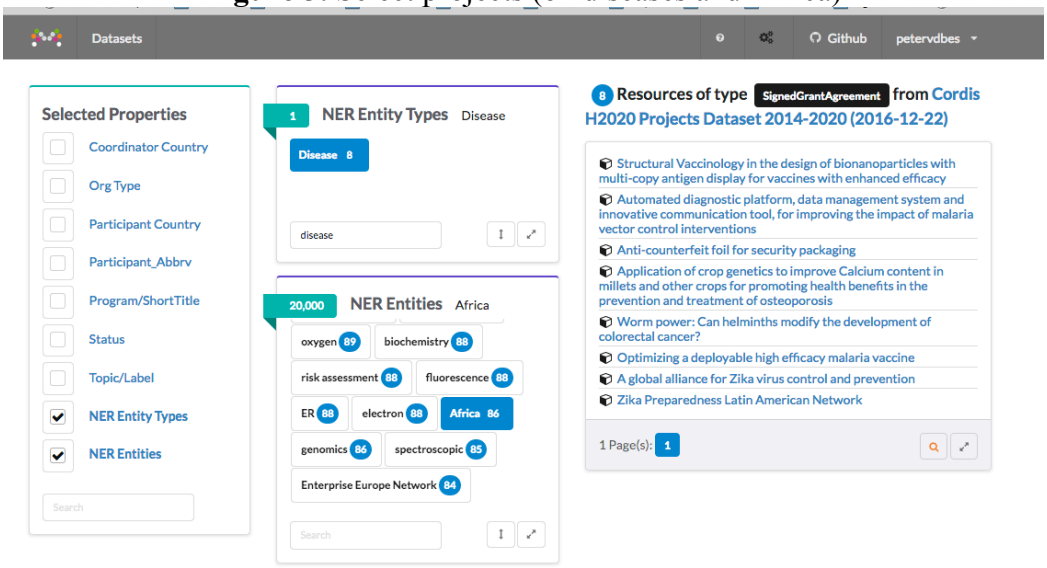
[6] http://www.dbpedia-spotlight.org/

[7] One may also add the *rejected* applications, to compare the portfolio of accepted applications with the rejected applications. This would help evaluating the selection process: are relevant topics systematically rejected?

We use as knowledge base the open DBpedia dataset, a standard in the open data community. This is the database under Wikipedia, in a useful (machine readable) format, and functions here as the knowledge base to annotate the project descriptions. The advantage is that the knowledge base is a product of communities improving its quality, and not the product of individual experts. As said we are planning to add specific field related taxonomies.

Named Entity Recognition: The SMS system has an 'entity recognition' functionality using DBpedia. Entities included in DBpedia are recognized in the project descriptions. This may need some pre-processing as the process is case sensitive. As DBpedia is a knowledge graph, the projects are linked to specific places in the knowledge graph, and though the graph systematically related to each other. In the current version of SMS, entities are partly subsumed under higher level Entity Types, and partly single Entities. We can use the knowledge graph to select projects. For example, by selecting a main category, e.g., *disease*, we only get projects that have in their description *a term referring to a disease* (figure 2). As these projects are also annotated with other terms, one may add a different dimension, e.g., the geographical dimension of diseases. By selecting within *diseases,* the category *continent* and within that *Africa*, we get all projects on diseases and (in) Africa (figure 3)

Figure 3: Select projects (on diseases and Africa)



This combining of terms has a great advantage, as we can combine *technical research* terms and *policy related terms* to retrieve the relevant projects. This may solve the problem of finding how research links to the grand societal challenges. This is a core problem in assessing relevance of research (described in technical terms and policy related terms).[8] Because the resulting set for a very specific topic is generally not too large, we can even manually inspect the policy-science link.

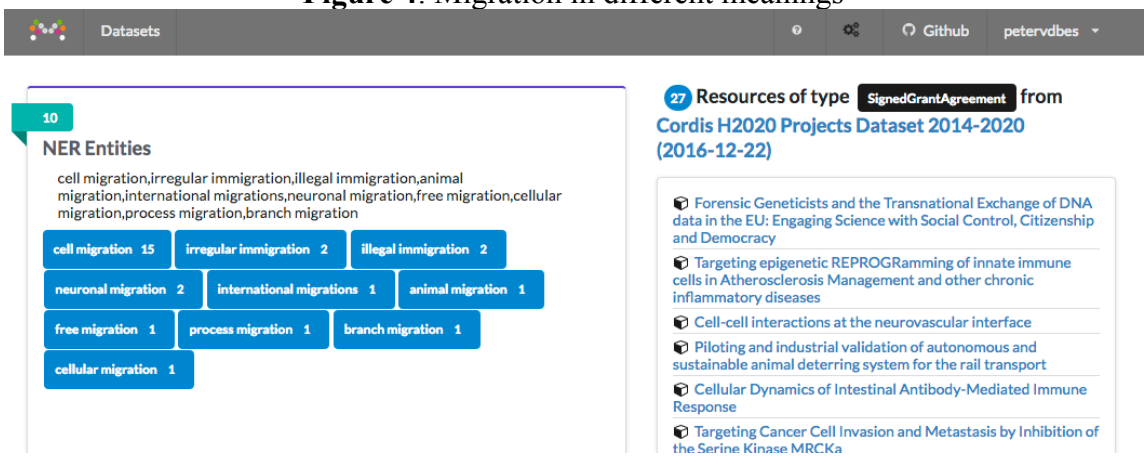
Now a crucial point is the quality of the knowledge bases, of the taxonomies. The larger the set of terms in the taxonomy, and the better the structure of cognitive links between the terms, the better the annotation works, and the better we can represent

Combining terms also helps to separate texts with similar terms. For example, migration appears in only a small number of H2020 projects. However, it appears in many different

[8] Peter van den Besselaar, *From priorities to projects to output*. Report for the Expert Group on Evaluation Methodologies. Amsterdam 2016a.

meanings, such as cell migration, neuronal migration, animal migration, branch migration, next to legal and illegal migration of people (Figure 4). One can then easily select the papers with the intended meaning of migration. This goes much quicker than we experienced when using searching the excel version of H2020.

Figure 4: Migration in different meanings



The other information available within Cordis makes more insight in the portfolio easy: after having selected a set of projects, we can easily find out see who is involved (organizations, countries), where the selected projects are located in the larger H2020 program (work program; sub-program), the funding level, etc. (figure 3).

Furthermore, we can use the advantage of a linked data approach: by linking the Cordis dataset with other data sets, we can find other properties of the organizations participating in the project. Linking to the geo-services enables to geo-locate the projects. Linking to the Web of Science may be used to find out what other specialties the project partners have, and how they collaborate. Linking the project partners to patent databases gives more information about the innovative activities, which the can be used for investigating a part of the impact.

EXAMPLE CASES

Chemistry for agriculture

We now look at chemical research in H2020 projects, related to one of the societal challenges. One could take e.g., *Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the Bio-economy*. This is rather broad, and therefore one may take smaller topics, such as agriculture, water, of sustainability. To investigate the portfolio, we have annotated all H2020 projects using DBpedia. We now first select all main entries (NER entries). In fact, there is one: *chemical substances*. We select this one (figure 5). We identify 976 projects that refer to *chemical substances* out of in total 11069 projects that up to now are funded in H2020. NER entity types are the main 251 categories, but there are many (20.000) detailed sub-categories: the NER entities. By clicking this in the faceted browser, we can see what other annotations these 976 projects did get: in total, some 7000. One can then relate the projects on chemical substances with other NER entities, for example *agriculture*. This is covered by three NER entities, and by selecting those in the browser, we see that this relates to only 19 granted projects, as figures 5 and 6 show. One may now easily browse through these projects to further find out where they are about.

Figure 5: Selecting the projects on chemical substances

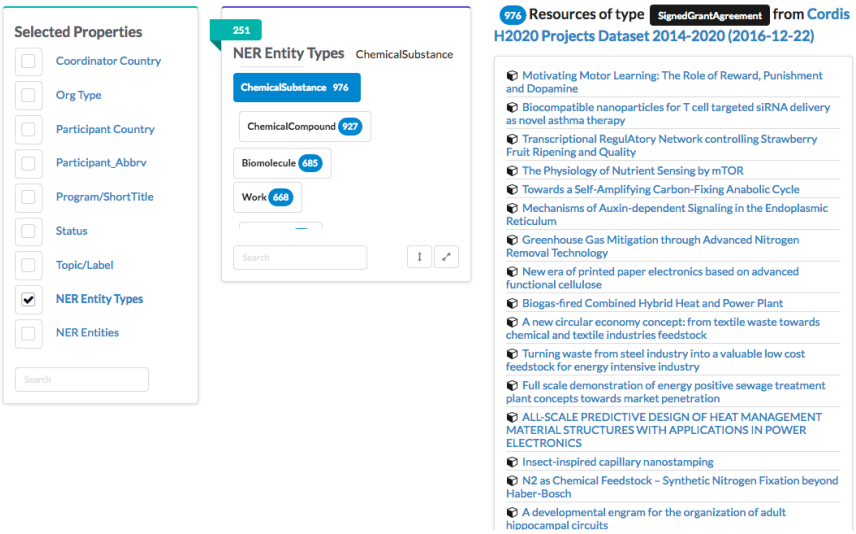
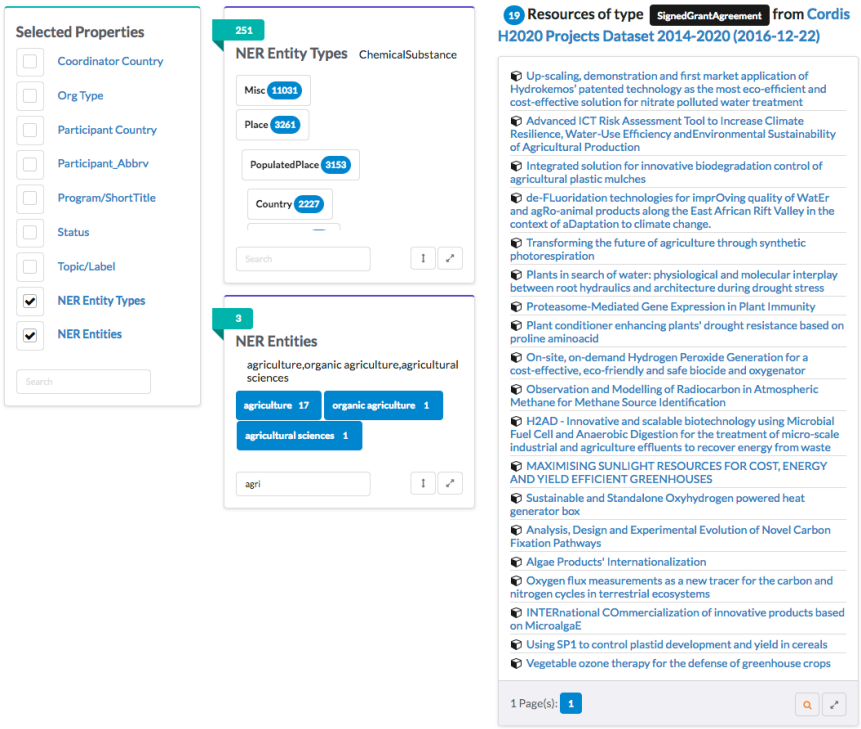


Figure 6: Identifying chemistry for agriculture



Water research

There are quite some water related topics in the H2020 projects, as figure 7 shows, and the 65 NER Entities identify some 200 projects. This can be easily refined to chemistry related project. Figure 8 shows the resulting list of 45 projects. Of the 65 water-related NER entities, some 25 NER entities are also related to chemistry, with about 45 projects. In total 22.5% of the water projects seem related to chemistry. Going a little deeper into this case may show the multidisciplinary character of the water related research in H2020, and what disciplines are more and what are less important in the portfolio.

Figure 7: Identifying water topics in H2020 – 200 projects

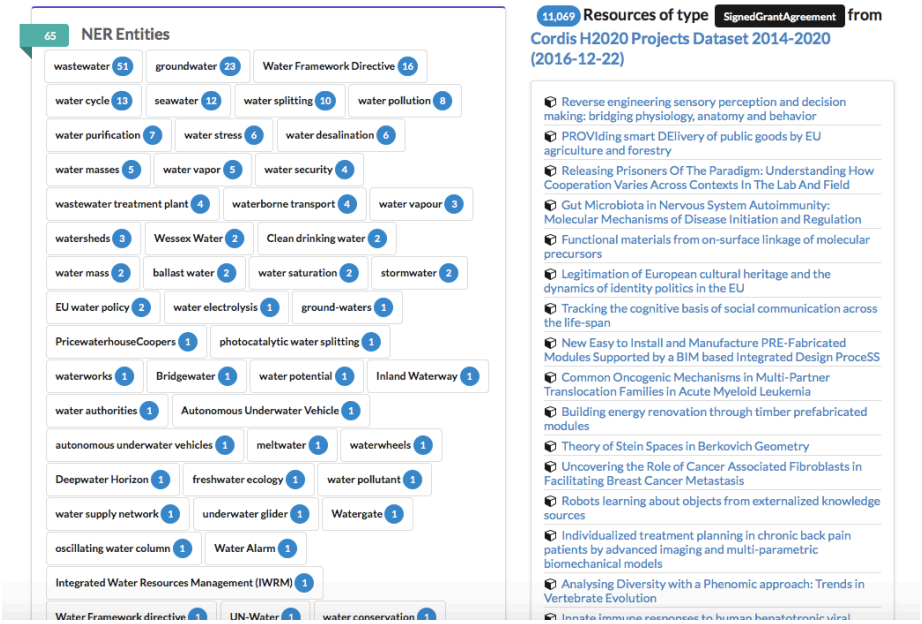
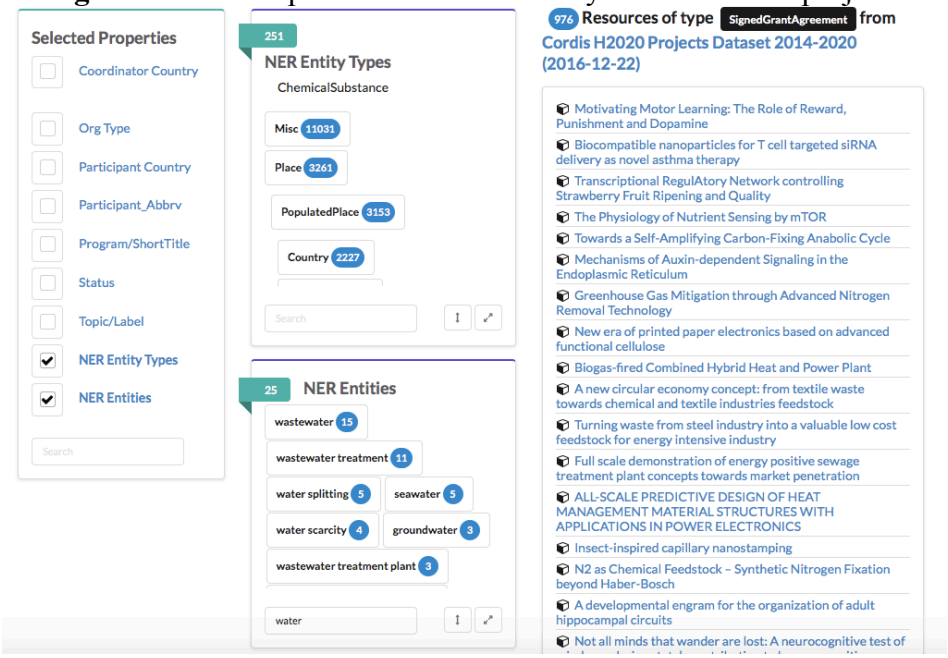
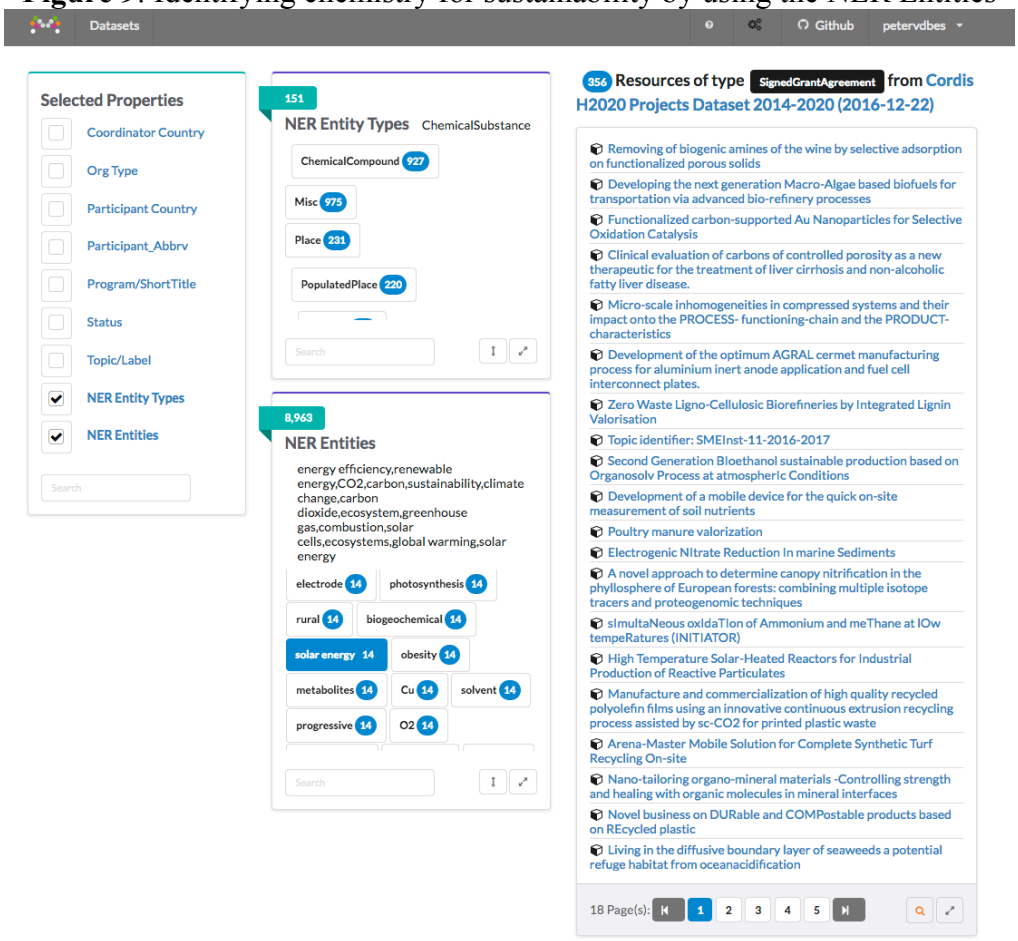


Figure 8: Water topics related to chemistry in H2020 – 45 projects



Chemistry for sustainability:
This is another 'cross section' of a research domain and a societal priority. We take the 976 chemistry projects as starting point and then select NER entities. The browser gives than all NER entities that are linked to these chemistry projects: 8963 NER Entities. They are listed from those that occur most often to those that occur only one time. As the NER Entities are no isolated terms but in a 'semantic hierarchy', it is smart to browse the NER Entities menu from top to bottom (figure 9).

Figure 9: Identifying chemistry for sustainability by using the NER Entities



When selecting, one quickly experiences that less frequent sub-categories are not adding any projects to the list, as they are already included in higher level categories. In the interface (Fig 9 – under NER entities), one can find which sub-categories are selected: Energy efficiency: Renewable energy; CO2; Carbon; Sustainability; climate change; Carbon dioxide; Greenhouse gas; Combustion; Solar cells; Ecosystems; Global warming; Solar energy.
The browser (top of the right-hand window) shows that about 40% of the chemistry projects (356) is focusing on sustainability, which can be further analyzed. This is done in figure 10, where we selected *Org(anization) Type*, and *Participant Country*. One can now starts to formulate questions on how portfolios are distributed over countries, and over types of organizations. And is this distribution related to the problems of states or regions?

For example, figure 11 shows that the chemistry for sustainability have more Higher Education institutions than companies as participants. We now can for example investigate whether this is uniform, or whether this is different in the different countries.
Figure 11 shows this for the Netherlands. Obviously, the dominant position of HEI is much less clear in the Netherlands. By comparing the NER Entities distribution between countries, one may be able to show differences in research activities between countries. The faceted browser produces on the background a *sparql query* (Figure 12 left side) which can be used to retrieve the selected data for further analysis (Figure 12 right side). This needs some editing and therefore some computer skills. We did this for the chemistry for sustainability portfolio, and then it is possible to use the existing tools for analysis and visualization to come to an assessment in terms of fields covered and societal issues addressed – and in terms of gaps in the portfolio.

Figure 10: Chemistry for sustainability by country and organization type

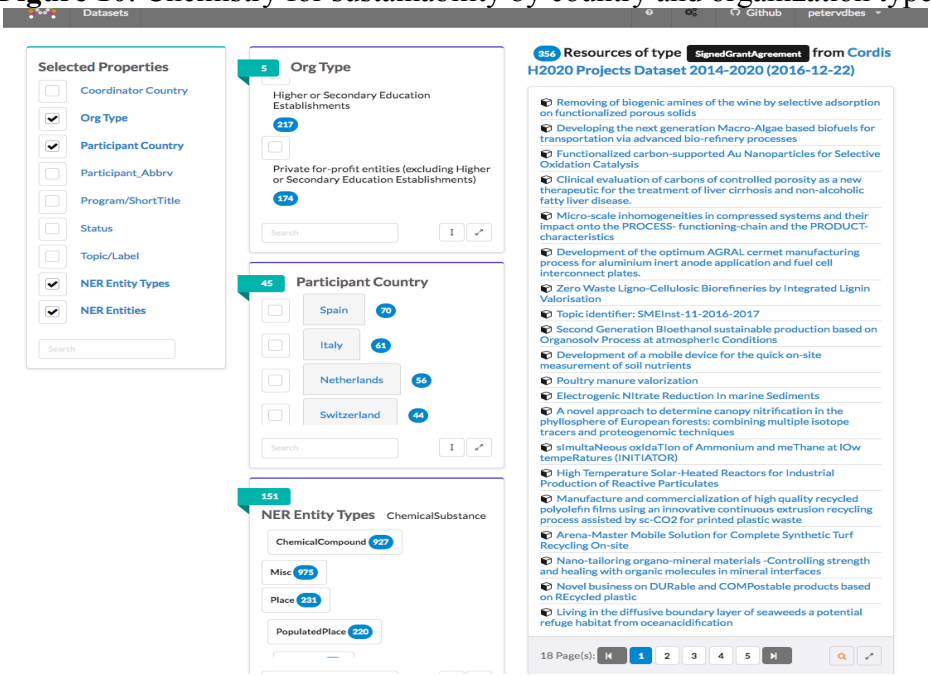
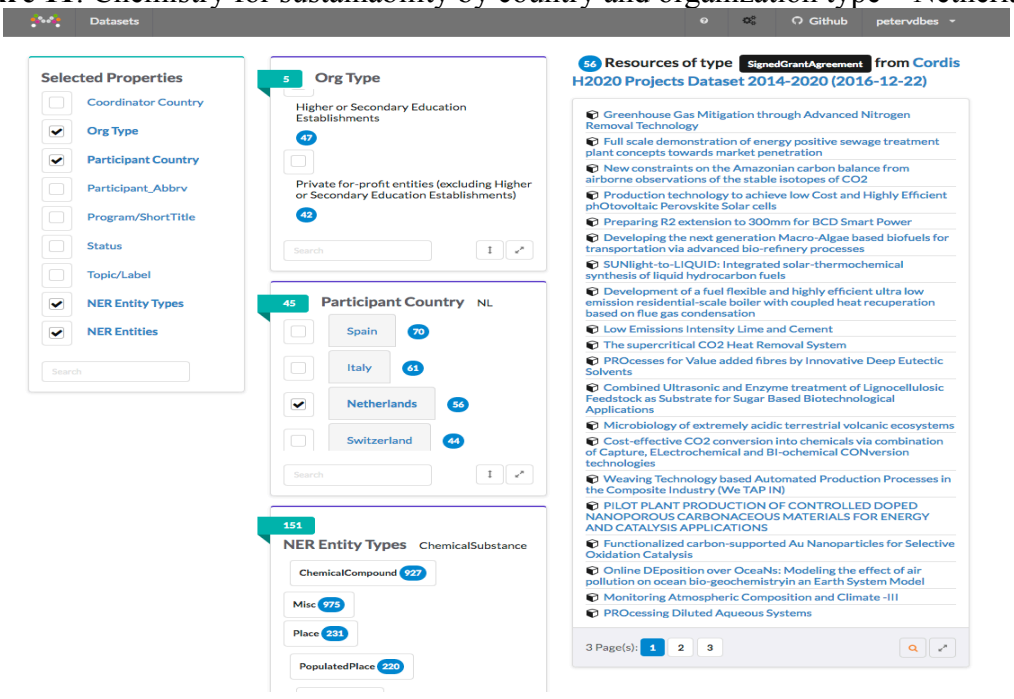


Figure 11: Chemistry for sustainability by country and organization type – Netherlands



CONCLUSIONS AND DISCUSSION

The procedure and tools shown in this paper enable the evaluation of a project portfolio (such as H2020) or an output portfolio (publications or patents of a research institution), in terms of its focus.

One may e.g., clarify the size of the part of a project portfolio that is mainly aimed at developing specific research fields (stimulating excellent research) and how big the part

devoted to specific societal challenges is (societal impact). As the research fronts and the societal challenges change over time, one may do the analysis for time slices of projects and evaluate the change of the quality of the portfolio over time: are the things addressed that one would want to? For example, in figure 4 we searched for all projects on 'migration', and show that only a few are related to todays increased migration flows from poor and war regions. Or one may evaluate the output of a research institute by annotating publications. What parts of the research front are covered, and where does the output relate to important societal questions. Are these overlapping sets of papers, of are these completely disjoint? And, is the output related to societal issues also of a good scholarly quality?

Figure 12: Chemistry for sustainability: query (partly) from the selection made in the browser, and resulting data table (partly)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX risis: <http://risis.eu/cordisH2020/vocab/>
PREFIX ldr: <https://github.com/alilk/ld-reactor/blob/master/vocab
SELECT DISTINCT ?projectID  ?orgAbbrv ?orgPIC ?orgCountry  WHERE {
  GRAPH <http://risis.eu/cordisH2020> {
    {
      SELECT DISTINCT ?s ?projectID ?objective ?orgAbbrv ?orgPIC ?
        GRAPH <http://risis.eu/cordisH2020> {
          ?s rdf:type risis:SignedGrantAgreement .
          ?s risis:id ?projectID .
          ?s risis:objective ?objective .
          ?s risis:participants ?participant .
          ?participant risis:abbreviation ?orgAbbrv .
          ?participant risis:id ?orgPIC .
          ?participant risis:country ?orgCountry .
          ?s ldr:annotations/ldr:surfaceForm ?v1.
          ?s ldr:annotations/ldr:uri/rdf:type ?v2.
          FILTER (str(?v1) IN ("""energy efficiency""","""renewabl
warming""","""carbon""","""sustainability""","""sustainable energy
gases""","""solar cell""","""ecosystem""","""ecosystems""","""comb
(<http://dbpedia.org/ontology/ChemicalSubstance>))
        }
      }
      ORDER BY ASC(?projectID)
    }

    OPTIONAL {
      ?s dcterms:title ?title .
    }
  }
}
```

Showing 1 to 50 of 1,702 entries (in 3.515 seconds)

	projectID	orgAbbrv	orgPIC	orgC
1	"633080"^^xsd:int	CNRS	"999997930"^^xsd:int	FR
2	"633080"^^xsd:int	CNRS Lyon	"999997930"^^xsd:int	FR
3	"633080"^^xsd:int	CNRS-IPSL	"999997930"^^xsd:int	FR
4	"633080"^^xsd:int	JRC	"999992304"^^xsd:int	BE
5	"633080"^^xsd:int	VUA	"954530344"^^xsd:int	NL
6	"633080"^^xsd:int	VU/VUmc	"954530344"^^xsd:int	NL
7	"633080"^^xsd:int	DLR	"999981731"^^xsd:int	DE
8	"633080"^^xsd:int	IPMA	"953379924"^^xsd:int	PT
9	"633080"^^xsd:int	AEMET	"996472271"^^xsd:int	ES
10	"633080"^^xsd:int	SRON	"997901663"^^xsd:int	NL
11	"633080"^^xsd:int	STICHTING SRON	"997901663"^^xsd:int	NL
12	"633080"^^xsd:int	DWD	"998059094"^^xsd:int	DE
13	"633080"^^xsd:int	EAA	"999452014"^^xsd:int	AT
14	"633080"^^xsd:int	SMHI	"999507983"^^xsd:int	SE
15	"633080"^^xsd:int	METEOROLOGISK INSTITUTT	"999510893"^^xsd:int	NO
16	"633080"^^xsd:int	KNMI	"999518944"^^xsd:int	NL
17	"633080"^^xsd:int	CERC	"999574428"^^xsd:int	UK
18	"633080"^^xsd:int	METEO-FRANCE	"999578890"^^xsd:int	FR
19	"633080"^^xsd:int	FMI	"999591306"^^xsd:int	FI
20	"633080"^^xsd:int	IASB - BIRA	"999642134"^^xsd:int	BE

FURTHER WORK

This paper demonstrates the potential for analyzing in detail the structure of a project or paper portfolio in terms of its scholarly and its societal profile. What would be the next steps?

- Inclusion of more knowledge graphs is needed: specialized ontologies/taxonomies representing research domains and societal challenges.
- Full text use for annotating, and testing parts of the text are important;
- Standard queries for retrieving parts of the portfolio for further inspection (e.g. using statistics or visualization); these standard queries would help the user without the computer skills needed to edit the automatically generated queries;
- When the dataset is very large, selecting in the browser takes time; further work on increasing the speed of querying in the browser would be useful.